Robust Bayesian Pitch Tracking Based on the Harmonic Model

Shi, L.; Nielsen, J. K.; Jensen, J. R.; Little, M. A.; Christensen, M. G.

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

DOI (link to publication from Publisher):
10.1109/TASLP.2019.2930917

Publication date:
2019

Document Version
Accepted author manuscript, peer reviewed version

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Robust Bayesian Pitch Tracking Based on the Harmonic Model

Liming Shi, *Student Member, IEEE,* Jesper Kjær Nielsen, *Member, IEEE,* Jesper Rindom Jensen, *Member, IEEE,* Max A. Little, and Mads Græsbøll Christensen, *Senior Member, IEEE*

***Abstract*—Fundamental frequency is one of the most important characteristics of speech and audio signals. Harmonic model-based fundamental frequency estimators offer a higher estimation accuracy and robustness against noise than the widely used autocorrelation-based methods. However, the traditional harmonic model-based estimators do not take the temporal smoothness of the fundamental frequency, the model order, and the voicing into account as they process each data segment independently. In this paper, a fully Bayesian fundamental frequency tracking algorithm based on the harmonic model and a first-order Markov process model is proposed. Smoothness priors are imposed on the fundamental frequencies, model orders, and voicing using first-order Markov process models. Using these Markov models, fundamental frequency estimation and voicing detection errors can be reduced. Using the harmonic model, the proposed fundamental frequency tracker has an improved robustness to noise. An analytical form of the likelihood function, which can be computed efficiently, is derived. Compared to the state-of-the-art neural network and non-parametric approaches, the proposed fundamental frequency tracking algorithm has superior performance in almost all investigated scenarios, especially in noisy conditions. For example, under 0 dB white Gaussian noise, the proposed algorithm reduces the mean absolute errors and gross errors by 15% and 20% on the Keele pitch database and 36% and 26% on sustained /a/ sounds from a database of Parkinson's disease voices. A MATLAB version of the proposed algorithm is made freely available for reproduction of the results[1].**

***Index Terms*—Fundamental frequency or pitch tracking, harmonic model, Markov process, harmonic order, voiced-unvoiced detection**

I. Introduction

THE problem of estimating the fundamental frequency or pitch information from noisy sound signals occurs in many applications, such as speech synthesis [1], voice disorder detection [2], and automatic speech recognition [3]. Fundamental frequency is a physical feature defined as the lowest frequency component of a periodic signal, while pitch is a perceptual feature, related to human listening [4]. Our objective is to estimate fundamental frequency. But, following [5], [6], we do not distinguish between fundamental frequency and pitch and use them interchangeably. Pitch is usually estimated using a segment of sound signals (a.k.a., frame) with a fixed segment length (e.g., 15-40 ms for speech signals [7]–[9]). Numerous pitch estimation algorithms have been proposed in the last fifty years, which can be categorized as unsupervised and supervised approaches. Unsupervised pitch estimation methods can be further categorized as non-parametric and parametric approaches. Examples of non-parametric approaches include the YIN [10], RAPT [11], SWIPE [12] and PEFAC [5] methods. YIN and RAPT compute autocorrelation functions from short frames of sound signals in the time domain. However, they are not robust against noise [13] and suffer from pitch octave errors (that is, a rational multiple of the true pitch) [3]. To reduce the pitch octave errors, SWIPE uses the cross-correlation function against a sawtooth signal combined with the spectrum of the signal, and exploits only the first and prime harmonics of the signal. PEFAC estimates the pitch in the log-frequency domain by convolving each frame's power spectrum with a filter that sums the energy of the pitch harmonics. Dynamic programming is used to obtain a smooth estimate of the pitch track. Due to the filtering and built-in spectral normalization methods, PEFAC is claimed to work in high levels of noise. However, a long frame length (e.g., 90.5 ms in PEFAC by default) is required to obtain good pitch estimation accuracy which is not practical in many real-time applications. More recently, a single frequency filtering approach based pitch estimation algorithm is proposed, which exploits the high SNR frequency component to overcome the effects of degradations in speech signal [14].

By contrast, parametric methods (e.g., harmonic model-based pitch estimators [6], [15], [16]) have also been proposed for pitch estimation. Compared with non-parametric approaches, harmonic model-based pitch estimators work with a short frame length (e.g., 20 ms), and show higher robustness to additive noise, fewer octave errors, and better time-frequency resolution [7], [17]. Recently, a computationally efficient pitch estimator based on a harmonic model has been proposed, which is referred to as the fast NLS [13]. However, one problem with most of the harmonic model based pitch estimators is that they do not take the temporal smoothness of the pitch, the harmonic order, and voicing into account as they process each frame independently. As a result, outliers, due to octave errors or voicing detection errors, occur. A sample-by-sample Kalman filtering-based pitch tracking algorithm using a time-varying harmonic model is proposed in [18] by assuming that the pitch and weights follow first-order Markov chains. A particle filtering-based pitch tracking algorithm based on

This work was funded by the Danish Council for Independent Research, grant ID: DFF 4184-00056.

L. Shi, J. K. Nielsen, J. R. Jensen and M. G. Christensen are with the Audio Analysis Lab, CREATE, Aalborg University, DK-9000 Aalborg, Denmark, e-mail: {ls, jrj, jkn,mgc}@create.aau.dk

M. A. Little is with the Engineering and Applied Science, Aston University and Media Lab, Massachusetts Institute of Technology, e-mail: max.little@aston.ac.uk

[1]An implementation of the proposed algorithm using MATLAB may be found in https://tinyurl.com/yxn4a543

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the source-filter speech model combining with the harmonic modelling of input source is introduced in [19]. However, the good performance of the algorithms in [18] and [19] requires careful initializations. Moreover, it is difficult to integrate the time-varying model order into these algorithms, see [20] as an example of combing discrete and continuous state spaces. With either a known or estimated model order, a maximum a posteriori (MAP) pitch estimator based on the harmonic model has been developed to exploit the temporal dynamics of the pitch [21]. The model weights and observation noise variance are estimated by maximizing the maximum likelihood function (i.e., a frequentist perspective). Smooth pitch estimates are obtained, and thus the pitch octave errors are reduced. An additional voicing state is further considered in [22] for estimating the pitch and obtaining the voiced-unvoiced decision jointly. However, the pitch tracking approach in [21] and [22] has many drawbacks. First, the assumption of a fixed harmonic order for multiple frames is not valid. In fact, in audio signals, the harmonic order often changes from frame to frame [9]. Second, matrix inversions are required to be stored for each candidate pitch to reduce the computational complexity. Third, errors can be found in transition frames where the voicing changes, because the past pitch information is not exploited when an unvoiced frame occurs. Finally, it is well-known that estimating parameters from a frequentist's perspective leads to over-fitting [23].

More recently, neural network based supervised pitch tracking algorithms were proposed [24]–[26], which show robustness against noise. The method proposed in [25] uses deep stacking network for joint speech separation and pitch estimation. The CREPE [26] discretises the pitch in logarithmic scale and uses a deep convolutional neural network to produce a pitch estimate. However, the unvoiced/silent state is not considered in the model. The maximum value of the output of the neural network is used as a heuristic estimate of the voicing probability. Moreover, to satisfy user's demand for different frequency resolution or frame length, the whole system is required to be retrained, which is usually time-consuming.

In this paper, we propose a fully Bayesian harmonic model-based pitch tracking approach. By using the harmonic model, as opposed to non-parametric methods, improved robustness against background noise and octave errors can be obtained. First-order Markov processes are used to capture the temporal dynamics of pitch, harmonic order, and voicing. By using information from previous frames, the rate of octave errors and the voicing detection errors can be further reduced. Compared to [21] and [22], we not only consider the temporal dynamics of pitch and voicing, but also of the harmonic order, which enables us to detect if any pitch is present, and estimate the pitch and harmonic order jointly and accurately. Moreover, past information on pitch is exploited to improve robustness against temporal voicing changes. Furthermore, by adopting a fully Bayesian approach to model weights and observation noise variances, the overfitting can be avoided. By assigning a proper transition pdf for the weights, fast NLS [13] can be easily incorporated into the proposed algorithm, leading to low computational and storage complexities.

The rest of the paper is organized as follows. In Section II, we briefly review general Bayesian tracking theory. In Section III and Section IV, we present the proposed harmonic observation and state evolution models, respectively. In Section V, the proposed pitch tracking algorithm is derived based on the harmonic observation and state evolution models. In Section VI, we briefly review the prewhitening step for dealing with non-Gaussian noise. Simulation results are given in Section VII, and the conclusions given in Section VIII.

Notation: Boldface symbols in lowercase and uppercase letters denote column vectors and matrices, respectively.

II. Bayesian tracking

In this section, we briefly review Bayesian tracking in general, which forms the fundamental structure of the proposed pitch tracking algorithm. Consider the problem of estimating the state sequence $\{\mathbf{x}_n\}, 1 \leq n \leq N$ from noisy observations $\{\mathbf{y}_n\}, 1 \leq n \leq N$, related by

$$\mathbf{y}_n = h(\mathbf{x}_n, \mathbf{v}_n), \tag{1}$$

where $h(\cdot)$ denotes a mapping function between the state and observation vectors, $\mathbf{v}_n$ denotes an i.i.d. observation noise sequence, and n denotes the time index. The state sequence follows a first-order Markov process:

$$\mathbf{x}_n = f(\mathbf{x}_{n-1}, \mathbf{m}_n), \tag{2}$$

where $f(\cdot)$ denotes a mapping function between the current and previous states, and $\mathbf{m}_n$ denotes an i.i.d. state noise sequence. The elements in the state vector $\mathbf{x}_n$ can either be continuous or discrete. Assume that the posterior pdf $p(\mathbf{x}_{n-1}|\mathbf{Y}_{n-1})$ is available with the initial pdf being defined as $p(\mathbf{x}_0)$, where $\mathbf{Y}_{n-1}$ denotes a collection of observation vectors from the first observation vector up to the $(n-1)^{\text{th}}$ observation vector, i.e.,

$$\mathbf{Y}_{n-1} = [\mathbf{y}_1, \cdots, \mathbf{y}_{n-1}].$$

The objective of Bayesian tracking is to obtain a posterior distribution over the state vector $\mathbf{x}_n$ based on the current and previous observations recursively, i.e., $p(\mathbf{x}_n|\mathbf{Y}_n)$. The posterior $p(\mathbf{x}_n|\mathbf{Y}_n)$ can be obtained in two stages: predict and update.

In the prediction stage, we obtain the prediction pdf $p(\mathbf{x}_n|\mathbf{Y}_{n-1})$ by using the transition pdf $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ from (2), i.e.,

$$\begin{aligned} &p(\mathbf{x}_n|\mathbf{Y}_{n-1}) \\ &= \int p(\mathbf{x}_n|\mathbf{x}_{n-1}) p(\mathbf{x}_{n-1}|\mathbf{Y}_{n-1}) d\mathbf{x}_{n-1}, \quad 2 \leq n \leq N, \\ &p(\mathbf{x}_1) = \int p(\mathbf{x}_1|\mathbf{x}_0) p(\mathbf{x}_0) d\mathbf{x}_0, \quad n = 1, \end{aligned} \tag{3}$$

which is known as the Chapman-Kolmogorov equation. Note that if the elements in $\mathbf{x}_n$ are all discrete variables, the integration operator should be replaced with the summation operator.

In the update stage, combining (1) and the prediction pdf from the prediction stage, Bayes' rule can be applied to obtain

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

the posterior, i.e.,

$$p(\mathbf{x}_n|\mathbf{Y}_n) = \frac{p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{Y}_{n-1})}{p(\mathbf{y}_n|\mathbf{Y}_{n-1})}, \quad 2 \leq n \leq N,$$
$$p(\mathbf{x}_1|\mathbf{Y}_1) = \frac{p(\mathbf{y}_1|\mathbf{x}_1)p(\mathbf{x}_1)}{p(\mathbf{y}_1)}, \quad n = 1, \tag{4}$$

where $p(\mathbf{y}_n|\mathbf{x}_n, \mathbf{Y}_{n-1})$ and $p(\mathbf{y}_1|\mathbf{x}_1)$ are the likelihood functions and $p(\mathbf{y}_n|\mathbf{Y}_{n-1})$ and $p(\mathbf{y}_1)$ are the normalization factors, respectively. Closed form solutions can be obtained for (3) and (4) in at least two cases. In the first case, when both $\mathbf{v}_n$ and $\mathbf{m}_n$ are drawn from Gaussian distributions with known parameters, and both $h(\mathbf{x}_n, \mathbf{v}_n)$ and $f(\mathbf{x}_{n-1}, \mathbf{m}_n)$ are linear functions over the variables, (3) and (4) reduce to the well-known Kalman-filter [23]. In the second case, when the state space is discrete and has a limited number of states, (3) and (4) reduce to the forward step of the forward-backward algorithm for hidden Markov model (HMM) inference [23]. In other cases, the inference of the posterior $p(\mathbf{x}_n|\mathbf{Y}_n)$ can be approximated using Monte Carlo approaches, such as particle filtering [27]. Next, we define the mapping function $h(\cdot)$ and formulate the observation equation (1) based on the harmonic model in Section III, and then explain the state evolution model (2) for the proposed pitch tracking algorithm in Section IV.

III. Harmonic observation model

A. The harmonic observation model

Consider the general signal observation model given by

$$\mathbf{y}_n = \mathbf{s}_n + \mathbf{v}_n, \tag{5}$$

where the observation vector $\mathbf{y}_n$ is a collection of M samples from the n^{th} frame defined as

$$\mathbf{y}_n = [y_{n,1}, \cdots, y_{n,M}]^T,$$

the clean signal vector $\mathbf{s}_n$ and noise vector $\mathbf{v}_n$ are defined similarly to $\mathbf{y}_n$, M is the frame length in samples and n is the frame index. We assume that $\mathbf{v}_n$ is a multivariate white noise processes with zero mean and diagonal covariance matrix $\sigma_n^2\mathbf{I}$, σ_n^2 is the noise variance, $\mathbf{I}$ is the identity matrix. When voiced speech or music is present, we assume that the pitch, model weights and model order are constant over a short frame (typically 15 to 40 ms for speech signals and longer for music signals) and $s_{n,m}$ (i.e., the m^{th} element of $\mathbf{s}_n$) follows the harmonic model, i.e.,

$$\mathrm{H}_1 : s_{n,m} = \sum_{k=1}^{K_n} \left[\alpha_{k,n}\cos(k\omega_n m) + \beta_{k,n}\sin(k\omega_n m)\right], \tag{6}$$

where $\alpha_{k,n}$ and $\beta_{k,n}$ are the linear weights of the k^{th} harmonic, $\omega_n = 2\pi f_n/f_{\text{s}}$ is the normalized digital radian frequency, f_{s} is the sampling rate, and K_n is the number of harmonics. When voiced speech/music is absent (unvoiced or silent), a null model is used, i.e.,

$$\mathrm{H}_0 : \mathbf{y}_n = \mathbf{v}_n. \tag{7}$$

Note that, based on the source-filtering model of speech generation, the unvoiced speech can be modelled as a coloured Gaussian process [28]. The observation noise in practice may have non-stationary and non-Gaussian properties, such as babble noise. However, we can deal with this by prewhitening the observation signals [9], which will be described in Section VI. Writing (6) in matrix form and combining (5) and (6) yields

$$\mathrm{H}_1 : \mathbf{y}_n = \mathbf{Z}(\omega_n, K_n)\mathbf{a}_{K_n} + \mathbf{v}_n, \tag{8}$$

where

$$\begin{aligned}
\mathbf{Z}(\omega_0, K_n) &= [\mathbf{c}(\omega_n), \cdots, \mathbf{c}(K_n\omega_n), \mathbf{d}(\omega_n), \cdots, \mathbf{d}(K_n\omega_n)],\\
\mathbf{c}(\omega_n) &= [\cos(\omega_n 1), \cdots, \cos(\omega_n M)]^T,\\
\mathbf{d}(\omega_n) &= [\sin(\omega_n 1), \cdots, \sin(\omega_n M)]^T,\\
\mathbf{a}_{K_n} &= [\alpha_{1,n}, \cdots, \alpha_{K_n,n}, \beta_{1,n}, \cdots, \beta_{K_n,n}]^T.
\end{aligned}$$

We can further write (7) and (8) together by introducing a binary voicing indicator variable u_n, i.e.,

$$\mathbf{y}_n = u_n\mathbf{Z}(\omega_n, K_n)\mathbf{a}_{K_n} + \mathbf{v}_n, \tag{9}$$

where $u_n \in \{0, 1\}$. When $u_n = 0$ and $u_n = 1$, (9) reduces to the unvoiced and voiced models (7) and (8), respectively.

We collect all the unknown variables into the state vector $\mathbf{x}_n = [\mathbf{a}_{K_n}, \sigma_n^2, \omega_n, K_n, u_n]^T$. Comparing (9) and (1), we can conclude that the mapping function $h(\cdot)$ is a nonlinear function w.r.t. the state vector $\mathbf{x}_n$. Moreover, the state vector $\mathbf{x}_n$ contains continuous variables $\mathbf{a}_{K_n}$, σ_n^2, ω_n and discrete variables K_n and u_n. However, due to the non-linear characteristics of (9) w.r.t. ω_n, uniform discretisation over the pitch ω_n is commonly used [13]. An off-grid estimate of ω_n can be obtained by pitch refinement algorithms, such as gradient descent [29]. Our target is to obtain estimates of the fundamental frequency ω_n, the harmonic order K_n, and the voicing indicator u_n, that is a subset of $\mathbf{x}_n$ defined as $\ddot{\mathbf{x}}_n = [\omega_n, K_n, u_n]^T$, from the noisy observation $\mathbf{y}_n$.

IV. The state evolution model

In this section, we derive the state evolution model (2) or more generally the transition probability density/mass function (pdf/pmf) $p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{Y}_{n-1})$ for continuous/discrete states of the proposed model. Following the fast NLS pitch estimation approach [13], we uniformly discretize the pitch $\omega_n \in \{\omega^f, 1 \leq f \leq F\}$ over the range $[\omega_{\min}, \omega_{\max}]$, where $\omega_{\min}$ and $\omega_{\max}$ denote the lowest and highest pitches in the searching space, respectively. Prior information can be used to set $\omega_{\min}$ and $\omega_{\max}$. For example, pitch is usually between 70 to 400 Hz for speech signals. The grid size is set to

$$\left\lfloor F\frac{\omega_{\max}}{2\pi} \right\rfloor - \left\lceil F\frac{\omega_{\min}}{2\pi} \right\rceil + 1,$$

where F denotes the DFT size for computing the likelihood function (see Section V and [13] for further details), $\lfloor\cdot\rfloor$ and $\lceil\cdot\rceil$ denote the flooring and ceiling operators, respectively. It is also shown that the optimal choice of F depends on the frame length and the harmonic order [13]. However, for simplicity and fast implementation, in this paper, we set $F = 2^{14}$. The state space for the discrete variables can be expressed as $\{\mathcal{M}(n) : [\omega_n = \omega^f, K_n = k, u_n = 1]^T, 1 \leq f \leq F, 1 \leq k \leq$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

$K^{\max}\} \cup \{\mathcal{M}_0(n) : u_n = 0\}$. The prediction pdf $p(\mathbf{x}_n|\mathbf{Y}_{n-1})$ defined in (3) can be factorized as

$$\begin{aligned} p(\mathbf{x}_n|\mathbf{Y}_{n-1}) =& p(\mathbf{a}_{K_n}|\sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \times \\ & p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) p(\ddot{\mathbf{x}}_n|\mathbf{Y}_{n-1}). \end{aligned} \tag{10}$$

We first explain the transition pdfs for the continuous variables σ_n^2 and $\mathbf{a}_{K_n}$, and then discuss the transition pmfs for the discrete variables ω_n, K_n and u_n. The selection of a state evolution model is a trade-off between being physically accurate and ending up with a practical solution.

A. Transition pdfs for the noise variance and weights

To obtain the prediction pdf for the noise variance $p(\sigma_n^2|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$, the transition pdf for the noise variance $p(\sigma_n^2|\sigma_{n-1}^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$ should be defined. A reasonable assumption for the noise variance is that it changes slowly from frame to frame. For example, the unknown parameter σ_n^2 can be assumed to evolve according to an inverse Gamma distribution [30], i.e.

$$p(\sigma_n^2|\sigma_{n-1}^2) = \mathcal{IG}(\sigma_n^2|c, d\sigma_{n-1}^2). \tag{11}$$

where $\mathcal{IG}(x|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-\alpha-1} \exp(-\frac{\beta}{x})$ and $\Gamma(\cdot)$ denotes the gamma function. With this transition pdf, an analytical form of the posterior distribution on $\mathbf{x}_n$ cannot be derived. A sequential Monte Carlo approach can be used to approximate the posterior numerically [31]. However, the major drawback of any Monte Carlo filtering strategy is that sampling in high-dimensional spaces can be inefficient [32]. A Rao-blackwellized particle filtering approach [33], which marginalises out some of the variables for statistical variance reduction, can be used to deal with this problem. However, we do not pursue this approach any further in this paper, and leave it for future work. Instead, for simplicity, we assume independence between σ_n^2 and σ_{n-1}^2, and use the Jeffery's prior, i.e.,

$$p(\sigma_n^2|\sigma_{n-1}^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) \propto 1/\sigma_n^2, \sigma_n^2 > 0. \tag{12}$$

As can be seen, the Jeffery's prior (12) is a limiting case of (11) with $c \to 0$ and $d \to 0$.

Similarly, we define the transition pdf for the weights as $p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}}, \sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$. Imposing smoothness dependency on the weight time evolution can reduce pitch octave errors [34]. However, in order to use the fast algorithm [13], we assume that the model weights between consecutive frames are conditionally independent given previous observations and the rest of unknown variables. Following [35], we use the hierarchical prior

$$\begin{aligned} & p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}}, \sigma_n^2, \ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}, g_n) \\ =& \mathcal{N}(\mathbf{a}_{K_n}|0, g_n\sigma_n^2 \left[(\mathbf{Z}(\omega_n, K_n)^T \mathbf{Z}(\omega_n, K_n)\right]^{-1}), \end{aligned} \tag{13}$$

$$p(g_n|\delta) = \frac{\delta - 2}{2}(1 + g_n)^{-\delta/2}, g > 0, \tag{14}$$

where $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that the vector $\mathbf{x}$ has the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The prior distribution for the weights (13) is known as Zellner's g-prior [36]. As can be seen from (13), given ω_n and K_n, the prior covariance matrix is a scaled version of the Fisher information matrix. With Zellner's g-prior, a closed-form calculation of the marginal likelihood can be obtained [37]. Moreover, the fast algorithm in [13] for computing the marginal likelihood can be applied (see Section V for detail).

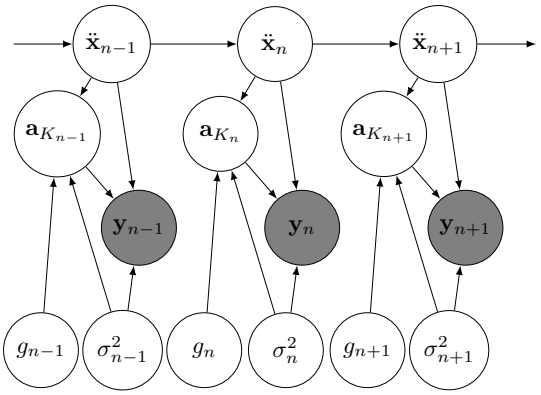

Fig. 1. A graphical model of the proposed method with shaded nodes indicating observed variables.

The graphical model for the proposed method is shown in Fig. 1. Note that, instead of obtaining point estimates of the noise variance and weight parameters using maximum likelihood [21], a Bayesian approach is used to represent the full uncertainty over these parameters.

B. Transition pmfs for ω_n, K_n and u_n

In [21], to reduce octave errors, a first-order Markov model is used for the pitch evolution provided that the harmonic order is fixed and known/estimated for multiple frames. Another voicing evolution model is further considered in [22] by imposing the so-called "hang-over" scheme [38]. Although in some cases, the harmonic order may not be of interest, it is still necessary to estimate it to obtain correct pitch estimates [39]. In fact, considering the temporal dynamics of the model order helps reducing the octave errors, which will be verified by the simulation results. Moreover, using priors for the model order is also necessary for model comparison [35]. In this paper, we propose to track the pitch ω_n, the harmonic order K_n and the voicing indicator u_n jointly. More specifically, we impose smoothness constraints on ω_n and K_n, and hang-over on voicing state using first-order Markov processes. The transition probability for the n^{th} frame to be voiced with pitch ω_n and harmonic order K_n when the previous frame is also voiced with ω_{n-1} and K_{n-1} can be expressed as

$$\begin{aligned} & p(\mathcal{M}(n)|\mathcal{M}(n-1)) \\ =& p(\omega_n, K_n|\omega_{n-1}, K_{n-1}, u_{n-1} = 1, u_n = 1) \times \\ & p(u_n = 1|u_{n-1} = 1). \end{aligned} \tag{15}$$

We assume that the pitch ω_n and harmonic order K_n evolve according to their own, independent dynamics given $u_n = 1$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

and $u_{n-1}=1$, i.e.,

$$\begin{aligned}&p(\omega_n,K_n|\omega_{n-1},K_{n-1},u_n=1,u_{n-1}=1)\\=&p(\omega_n|\omega_{n-1},u_n=1,u_{n-1}=1)\times\\&p(K_n|K_{n-1},u_n=1,u_{n-1}=1),\end{aligned}\tag{16}$$

which means when both time frame $n-1$ and n are voiced, the pitch and harmonic order only depend on their previous states. In fact, this assumption is only true when the product of the maximum allowed harmonic order and the pitch is less than half of the sampling frequency. However, by using a Bayesian approach, a model with a larger harmonic order is more penalized than with a smaller one. Even if a large value is used for the maximum allowed harmonic order in the proposed approach, the posterior model probability with a large harmonic order can be small [40]. In [41], an infinite number of harmonics is used, and the non-parametric prior distribution is used to penalize the models with large harmonic orders. By assuming the pitch and harmonic order are conditionally independent given $u_n=1$ and $u_{n-1}=1$, the Bayesian inference of the model posterior, shown in Section V, can be simplified. The transition probability for the n^{th} frame to be voiced with pitch ω_n and harmonic order K_n when the previous frame is unvoiced/silent can be expressed as

$$\begin{aligned}&p(\mathcal{M}(n)|\mathcal{M}_0(n-1))\\=&p(\omega_n,K_n|u_n=1,u_{n-1}=0)p(u_n=1|u_{n-1}=0).\end{aligned}\tag{17}$$

The priors from an unvoiced frame to a voiced frame $p(\omega_n,K_n|u_n=1,u_{n-1}=0)$ are set to $p(\omega_m,K_m|\mathbf{Y}_m,u_m=1)$, which can be calculated as

$$p(\omega_m,K_m|\mathbf{Y}_m,u_m=1)=\frac{p(\omega_m,K_m,u_m=1|\mathbf{Y}_m)}{1-p(u_m=0|\mathbf{Y}_m)},\tag{18}$$

where m is the closest frame index to n that satisfies the constraint $p(u_m=0|\mathbf{Y}_m)<0.5$ (m^{th} frame is voiced). In fact, if the previous frame is not voiced, we exploit the information from the latest frame that is voiced as the prior for the pitches and harmonic orders. The motivation for this choice is that the pitch and harmonic order usually do not change abruptly after a short segment of unvoiced/silent frames. Using the past information as the prior, robustness against the voicing state changes can be improved. The graphical model for the evolution of $\ddot{\mathbf{x}}(n)$ is shown in Fig. 2. Assuming the Markov processes are time-invariant, we can express the transition matrices for the pitch, harmonic order and voicing as $\mathbf{A}^{\omega}$, $\mathbf{A}^{K}$ and $\mathbf{A}^{u}$, respectively.

V. Pitch tracking

In this section, a joint pitch and harmonic order tracking, and voicing detection algorithm is derived based on the Bayesian tracking formulas (3) and (4). First, note that, by assuming that σ_n^2 and σ_{n-1}^2 are conditionally independent given $\ddot{\mathbf{x}}_n$ and $\mathbf{Y}_{n-1}$, and $\mathbf{a}_{K_n}$ and $\mathbf{a}_{K_{n-1}}$ are conditionally independent given σ_n^2, $\ddot{\mathbf{x}}_n$ and $\mathbf{Y}_{n-1}$, the prediction pdfs are

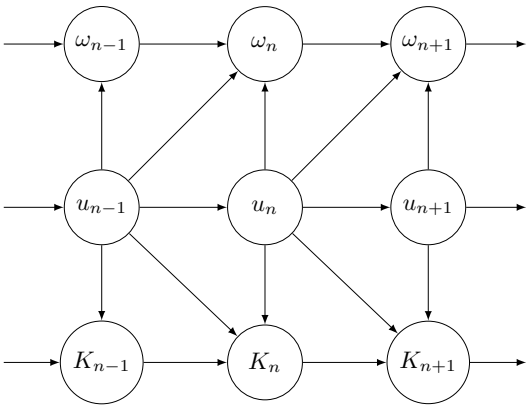

Fig. 2. A graphical model specifying conditionally independence relations for the discrete variables.

equal to the transition pdfs, i.e.,

$$p(\sigma_n^2|\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})=p(\sigma_n^2|\sigma_{n-1}^2,\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1}),\tag{19}$$

$$\begin{aligned}&p(\mathbf{a}_{K_n}|\sigma_n^2,\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})\\=&\int p(\mathbf{a}_{K_n}|\mathbf{a}_{K_{n-1}},\sigma_n^2,\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1},g_n)p(g_n;\delta)dg_n.\end{aligned}\tag{20}$$

Based on (3), prediction pmfs for discrete variables $p(\ddot{\mathbf{x}}(n)|\mathbf{Y}_{n-1})$ can be expressed as

$$\begin{aligned}&p(\mathcal{M}(n)|\mathbf{Y}_{n-1})\\=&\sum_{\mathcal{M}(n-1)}p(\mathcal{M}(n)|\mathcal{M}(n-1))p(\mathcal{M}(n-1)|\mathbf{Y}_{n-1})+\\&p(\mathcal{M}(n)|\mathcal{M}_0(n-1))p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1}),\end{aligned}\tag{21}$$

$$\begin{aligned}&p(\mathcal{M}_0(n)|\mathbf{Y}_{n-1})\\=&\sum_{h=0}^{1}p(u_n=0|u_{n-1}=h)p(u_{n-1}=h|\mathbf{Y}_{n-1})\\=&p(u_n=0|u_{n-1}=0)p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1})+\\&p(u_n=0|u_{n-1}=1)(1-p(\mathcal{M}_0(n-1)|\mathbf{Y}_{n-1})).\end{aligned}\tag{22}$$

With the prediction pdfs and pmfs in hand, we can obtain the update equation based on (4). In order to obtain the posteriors for the pitch, harmonic order and voicing indicators, the weights and noise variance can be integrated out from the update equation, i.e.,

$$\begin{aligned}&p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n)\\\propto&\int p(\mathbf{y}_n|\mathbf{x}_n,\mathbf{Y}_{n-1})p(\mathbf{x}_n|\mathbf{Y}_{n-1})d\mathbf{a}_{K_n}d\sigma_n^2\\=&p(\mathbf{y}_n|\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})p(\ddot{\mathbf{x}}_n|\mathbf{Y}_{n-1}),\end{aligned}\tag{23}$$

where $p(\mathbf{y}_n|\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})$ denotes a marginal likelihood, defined as

$$\begin{aligned}p(\mathbf{y}_n|\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})=\int p(\mathbf{y}_n|\mathbf{x}_n)p(\mathbf{a}_{K_n}|\sigma_n^2,\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})\times\\p(\sigma_n^2|\ddot{\mathbf{x}}_n,\mathbf{Y}_{n-1})p(g_n;\delta)d\mathbf{a}_{K_n}d\sigma_n^2dg_n.\end{aligned}\tag{24}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

Using (9), (12), (13), (14), (19) and (20), a closed-form marginal likelihood can be obtained, i.e.,

$$p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1}) = \left[\frac{(\delta-2)}{2K_n+\delta-2}{}_2F_1\left[\frac{M}{2}, 1; \frac{2K_n+\delta}{2}; R^2(\omega_n, K_n)\right]\right]^{u_n} \times m_M(\mathbf{y}_n), \quad (25)$$

where

$$m_M(\mathbf{y}_n) = \frac{\Gamma(\frac{M}{2})}{(\pi||\mathbf{y}_n||_2^2)^{\frac{M}{2}}}, \quad (26)$$

$$R^2(\omega_n, K_n) = \frac{\mathbf{y}_n^T\mathbf{Z}(\omega_n, K_n)\hat{\mathbf{a}}_{K_n}}{\mathbf{y}_n^T\mathbf{y}_n}, \quad (27)$$

$$\hat{\mathbf{a}}_{K_n} = (\mathbf{Z}(\omega_n, K_n)^T\mathbf{Z}(\omega_n, K_n))^{-1}\mathbf{Z}(\omega_n, K_n)\mathbf{y}_n, \quad (28)$$

$m_M(\mathbf{y}_n)$ denotes the null model likelihood (i.e., $p(\mathbf{y}_n|u_n = 0)$) and ${}_2F_1$ denotes the Gaussian hypergeometric function [42]. To compute $R^2(\omega_n, K_n)$ for all the candidate pitches and harmonic orders, the fast algorithm [13] can be applied. Moreover, from a computational point of view, a Laplace approximation of (24) can be derived as an alternative instead of marginalizing w.r.t. g_n analytically [35]. Note that, for the discrete vector $\ddot{\mathbf{x}}_n$, it should satisfy the normalisation constraint,

$$1 = \sum_{\ddot{\mathbf{x}}_n} p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n) = p(\mathcal{M}_0(n)|\mathbf{Y}_n) + \sum_{\mathcal{M}(n)} p(\mathcal{M}(n)|\mathbf{Y}_n). \quad (29)$$

Finally, estimates of the pitch and harmonic order and the voiced/unvoiced state can be jointly obtained using the maximum a posterior (MAP) estimator. More specifically, the n^{th} frame is labeled as voiced if $p(u_n = 0|\mathbf{Y}_n) < 0.5$, and the pitch and harmonic order are obtained as

$$(\hat{\omega}_n, \hat{K}_n) = \max_{\omega_n, K_n} p(\omega_n, K_n, u_n = 1|\mathbf{Y}_n). \quad (30)$$

The proposed Bayesian pitch tracking algorithm is shown in Algorithm 1. To make inferences, we need to specify the transition matrices for the pitch $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1)$, the harmonic order $p(K_n|K_{n-1}, u_n = 1, u_{n-1} = 1)$ and $p(u_n|u_{n-1})$. Following [21], we set $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1) = \mathcal{N}(\omega_n|\omega_{n-1}, \sigma_\omega^2)$. The transition probability for the model order is chosen as $p(K_n|K_{n-1}, u_n = 1, u_{n-1} = 1) = \mathcal{N}(K_n|K_{n-1}, \sigma_K^2)$. Smaller σ_ω^2 and σ_K^2 lead to smoother estimates of the pitch and harmonic order while larger values make the inference less dependent on the previous estimates. The matrix $\mathbf{A}^u$ is controlled by $p(u_n = 1|u_{n-1} = 0)$ and $p(u_n = 0|u_{n-1} = 1)$. In order to reduce the false negative (wrongly classified as unvoiced when a frame is voiced) rate, we set $p(u_n = 1|u_{n-1} = 0) = 0.4$, $p(u_n = 0|u_{n-1} = 1) = 0.3$, respectively, that is, the transition probability from unvoiced to voiced is higher than from voiced to unvoiced. Note that each row of $\mathbf{A}^\omega$, $\mathbf{A}^K$, and $\mathbf{A}^u$ is normalised to ensure they are proper pmfs. By setting $\sigma_\omega^2 \to \infty$, $\sigma_K^2 \to \infty$, $p(u_n = 1|u_{n-1} = 0) = 0.5$ and $p(u_n = 0|u_{n-1} = 1) = 0.5$, the proposed algorithm reduces to the fast NLS algorithm [13]. Moreover, using (16), (18), and the definitions of $\mathbf{A}^\omega$, $\mathbf{A}^K$ and $\mathbf{A}^u$, an MAP estimator that maximizes the joint posterior $p(\ddot{\mathbf{x}}_1, \cdots, \ddot{\mathbf{x}}_N|\mathbf{Y}_N)$, instead of marginal posterior $p(\ddot{\mathbf{x}}_n|\mathbf{Y}_n)$ in (23), can also be derived, which is known as the Viterbi algorithm [23]. Although the Viterbi algorithm may help obtaining better pitch estimates by using future data, it has high storage complexity. In this paper, we only focus on the online pitch tracking in Algorithm 1.

Algorithm 1 The proposed Bayesian pitch tracking

Initiate the harmonic order $K^{\max}$, transition matrices $\mathbf{A}^\omega$, $\mathbf{A}^K$ and $\mathbf{A}^u$, and the initial probability $p(u_0|\mathbf{y}_0)$ and $p(\omega_0, K_0, u_0 = 1|\mathbf{y}_0)$ satisfying the constraint $p(u_0 = 0|\mathbf{y}_0) + \sum_{\omega_0, K_0} p(\omega_0, K_0, u_0 = 1|\mathbf{y}_0) = 1$
for $n = 1, 2, \cdots$ **do**
  Prediction step:
  Obtain $p(\mathcal{M}(n)|\mathbf{Y}_{n-1})$ based on (21), (15) and (17).
  Obtain $p(\mathcal{M}_0(n)|\mathbf{Y}_{n-1})$ based on (22).
  Update step:
  Calculate $p(\mathbf{y}_n|\ddot{\mathbf{x}}_n, \mathbf{Y}_{n-1})$ using the fast weight estimation algorithm [13] and (25).
  Calculate the unnormalised posteriors $p(\mathcal{M}(n)|\mathbf{Y}_n)$ and $p(\mathcal{M}_0(n)|\mathbf{Y}_n)$ based on (23).
  Normalise the posteriors based on the constraint (29).
  MAP estimation:
  if $p(\mathcal{M}_0(n)|\mathbf{Y}_n) > 0.5$ **then**
    The n^{th} frame is labeled as unvoiced/silent.
  else
    The n^{th} frame is labeled as voiced.
    Estimating $\hat{\omega}_n$ and $\hat{K}_n$ based on (30).
    Update $p(\omega_m, K_m|\mathbf{Y}_m, u_m = 1)$ based on (18).
  end if
end for

VI. Prewhitening

The fast NLS and proposed pitch tracking algorithm are derived under the assumption of white Gaussian noise. However, this assumption is usually violated in practice, for example, babble noise in a conference hall. Therefore, a prewhitening step is required to deal with the inconsistency between the white Gaussian noise model assumption and real life noise model. A linear prediction (LP) based prewhitening step is applied to each frame to deal with the non-white Gaussian noise (see [9], [43] for detail). The power spectral density (PSD) of the noise given noisy signals is estimated using the minimum mean-square error (MMSE) estimator [44]. We refer to the fast NLS and proposed algorithm with prewhitening step as Prew-Fast NLS and Prew-Proposed, respectively. Combing the prewhitening step and Algorithm 1, a block diagram for the proposed pitch tracker in colored noise scenarios is shown in Fig. 3, where $\mathbf{y}_n^{prew}$ denotes the prewhitened observation vector and $\hat{\ddot{\mathbf{x}}}_{n-1}$ denotes an estimate of $[\omega_n, K_n, u_n]^T$.

VII. Simulation

In this section, we test the performance of the proposed harmonic model-based pitch tracking algorithm on real speech signals.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

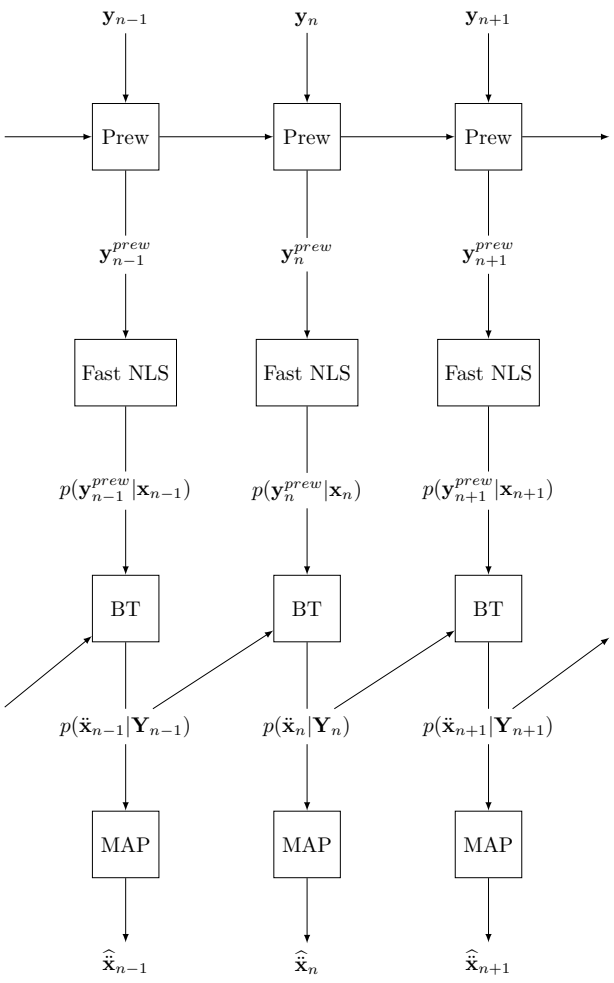

Fig. 3. A block diagram of the proposed algorithm with prewhitening for colored noise, where Prew, and BT are abbreviations for prewhitening, Bayesian tracking, respectively.

A. Databases

The databases used for evaluating the performance of different algorithms are as follows:

MIS database: containing 300 audio files from 6 different instrument classes: piano, violin, cello, flute, bassoon, and soprano saxophone, at a sampling rate of 44.1 kHz[2].

Keele pitch database: containing 10 spoken sentences from five male and five female speakers at a sampling rate of 20 kHz [45]. The "ground truth" pitch estimates are extracted from electroglottography with 10 ms time frame increment and 25.6 ms frame length. In fact, there are many spikes and wrong estimates in the "ground truth" pitch values, especially in the transient frames. However, we present the results for the Keele database to facilitate comparison with other pitch estimation algorithms that use this database.

Parkinson's disease database: containing 130 sustained /a/ phonations from patients with Parkinson's disease [46] at a sampling rate of 44.1 kHz. Each of the phonations is in one second length. The estimated "ground truth" pitches in 10 ms time frame increment are extracted from electroglottography (EGG).

[2]Audio files available in http://theremin.music.uiowa.edu

B. Performance measures

Three performance measures are considered:

Total error ratio (TER) [22]: voicing detection performance measure. It is calculated based on the ratio between the number of incorrect voicing detection (false positive and true negative) estimates and the number of total estimates.

Gross error ratio (GER) [12]: accuracy measure of pitch estimates. It is computed based on the ratio between the number of pitch estimates that differ by more than 20 percents from the ground truth and the number of total estimates. The unvoiced frames from the ground truth are excluded and the pitch value of the voiced frame that is wrongly labeled as unvoiced frames by different pitch estimation algorithms is set to 0.

Mean absolute error (MAE) [46]: accuracy measure of pitch estimates. It is computed based on mean of the absolute errors between the ground truth and estimates. The unvoiced frames from the ground truth are excluded and the oracle voicing detector is used for all the algorithms.

C. Experimental results on speech and audio samples

In this subsection, the experimental results of different pitch estimation algorithms for one speech and one audio sample, are presented in the first and second experiments, respectively.

First, the proposed approach is tested on concatenated speech signals uttered by a female speaker first, male speaker second, sampled at 16 kHz [3]. The spectrogram of the clean speech signals, pitch estimates, order estimates and the voicing detection results for PEFAC, CREPE, YIN, fast NLS and the proposed algorithm are shown in Fig. 4. The time frames of the spectrograms without red lines on top are unvoiced or silent frames. The variances for the transition matrices σ_ω^2 and σ_K^2 are set to $\frac{16\pi^2}{f_s^2}$ and 1, respectively. The SNR for white Gaussian noise is set to 0 dB. The candidate pitch ω_0 is constrained to the range $2\pi\,[70\ 400]\,/f_s$ for PEFAC, YIN, fast NLS and the proposed algorithm. However, the results for the neural network based approach CREPE is based on the model with the pitch range $2\pi\,[32.7\ 1975.5]\,/f_s$ provided by the authors [26]. To change the settings for CREPE, re-training of the neural network model is required. The maximum allowed harmonic order for the proposed and fast NLS is set to 10. The frame length is $M = 400$ samples (25 ms) with 60% overlap (10 ms time frame increment). As can be seen from Fig. 4, the voicing detection results of both the fast NLS and the proposed algorithm are better than those of YIN, PEFAC and CREPE. For example, the frames around 2.8 s are correctly classified as voiced by the fast NLS and the proposed, but wrongly labeled as unvoiced by YIN, PEFAC and CREPE. Fast NLS suffers from octave errors, and has outliers particularly in the transition frames where voicing decisions change. In the transition frame around 1.8 s, the pitch and number of harmonics are wrongly estimated to

[3]The example speech signal file is available in https://tinyurl.com/yxn4a543

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

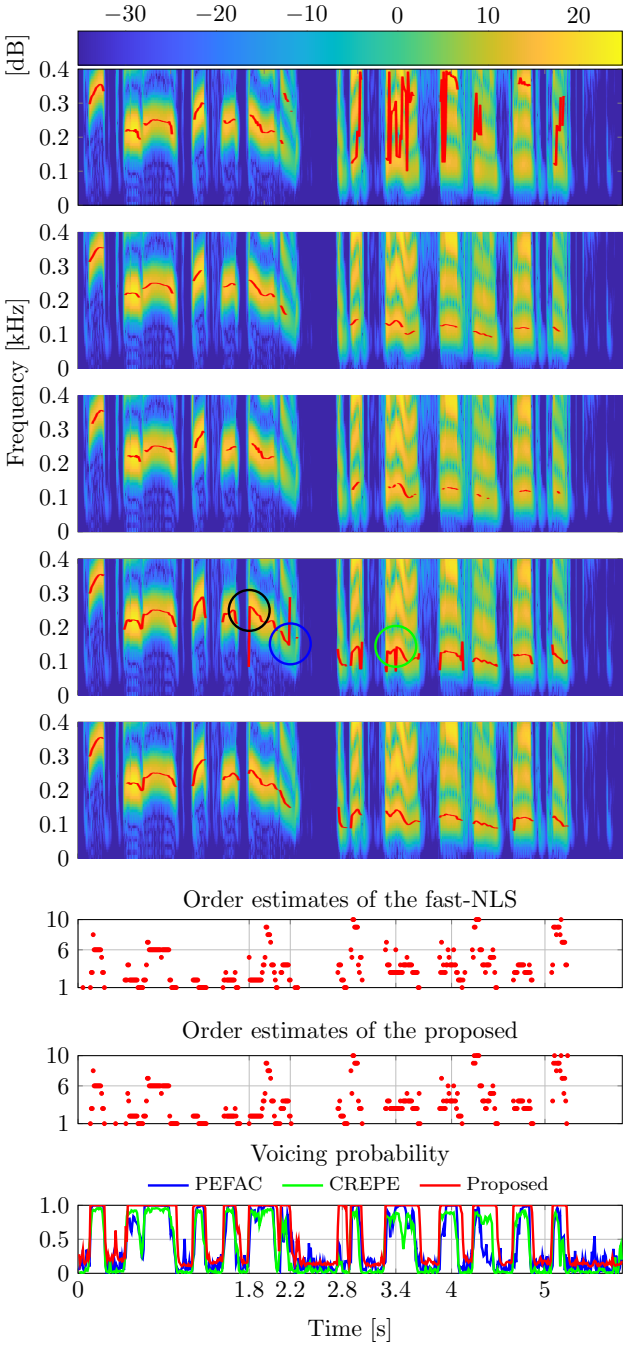

Fig. 4. Pitch estimates from PEFAC, CREPE, YIN, fast NLS and the proposed, the order estimates of the fast NLS and the proposed, and the voicing probabilities for real speech signals in 0 dB white Gaussian noise (from top to bottom).

84.8 Hz and five, respectively, by the fast NLS. In contrast, they are estimated to 248.8 Hz and one, respectively, by the proposed. Clearly, the estimates of the proposed fit better to the spectrogram than the estimates of the fast NLS. The reason for the robustness against transient frames of the proposed algorithm is that the pitch and harmonic order information of the latest voiced frame is used as the prior, i.e. (18). The harmonic order of the frame in 2.2 s is estimated to two by both the fast NLS and the proposed. However, the

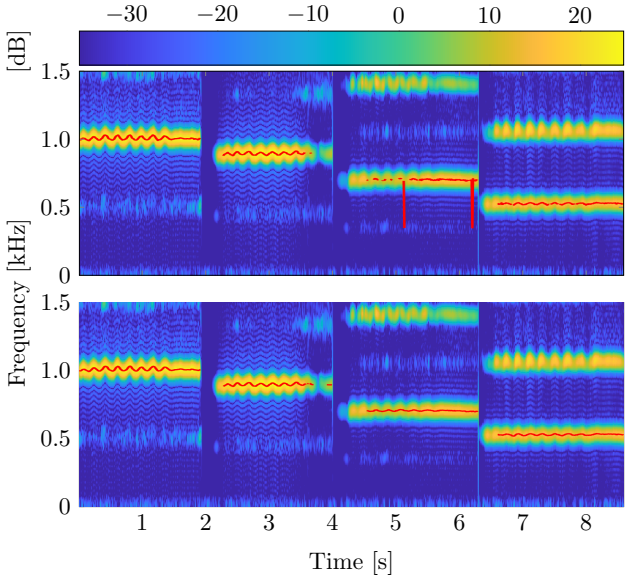

Fig. 5. Pitch estimates of fast NLS and the proposed algorithm for musical sounds in -5 dB white Gaussian noise (from top to bottom).

pitch is wrongly estimated to 288.8 Hz by the fast NLS, but correctly estimated to 150.4 Hz by the proposed. By imposing temporal smoothness prior on the pitch using the Markov process model $p(\omega_n|\omega_{n-1}, u_n = 1, u_{n-1} = 1)$, smoother estimates of the pitches are obtained. An octave error is produced by the fast NLS in the frame around 3.4 s. The pitch and harmonic order are estimated to 72 and six, respectively, by the fast NLS, but 143.2 and three, respectively, by the proposed. In fact, harmonic orders are estimated to three in the surrounding frames by both the fast NLS and the proposed. By using Bayesian tracking for the pitches and harmonic orders, smoother estimates of the pitches and harmonic orders are obtained. In conclusion, the proposed Bayesian pitch tracking algorithm obtains smooth estimates of the pitch and harmonic orders, and good voicing detection results by exploiting the past information.

The second experiment tests the performance of the proposed algorithm on musical instrument sounds (flute) from MIS database, decreasing from note B5 to C5. The spectrogram of the clean signals and the pitch estimates from fast NLS and the proposed algorithm are shown in Fig. 5. The music signal is downsampled to 16 kHz. The SNR for Gaussian white noise is set to -5 dB. The pitch ω_0 is constrained to the range $2\pi\,[100\ 1500]\,/f_s$. The other parameters are set to the same as for Fig. 4. As can be seen, the proposed algorithm not only has smoother estimates of the pitch than fast NLS but also better voicing detection results.

D. Experimental results on the Keele pitch database

In this subsection, the experimental results of different pitch estimation algorithms, using the Keele database, in white Gaussian noise, colored noise and reverberated conditions are presented.

First, we test the performance of the proposed algorithm on the Keele pitch database with white Gaussian noise. TER,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

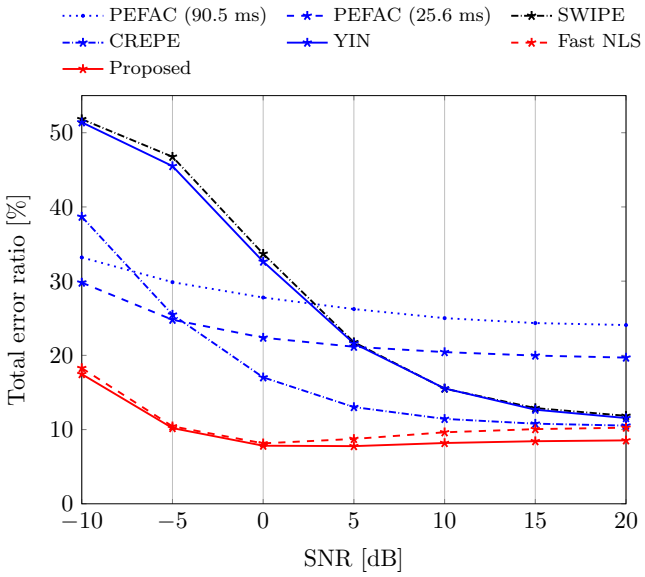

Fig. 6. Total error ratio in different SNRs for the Keele pitch database in white Gaussian noise

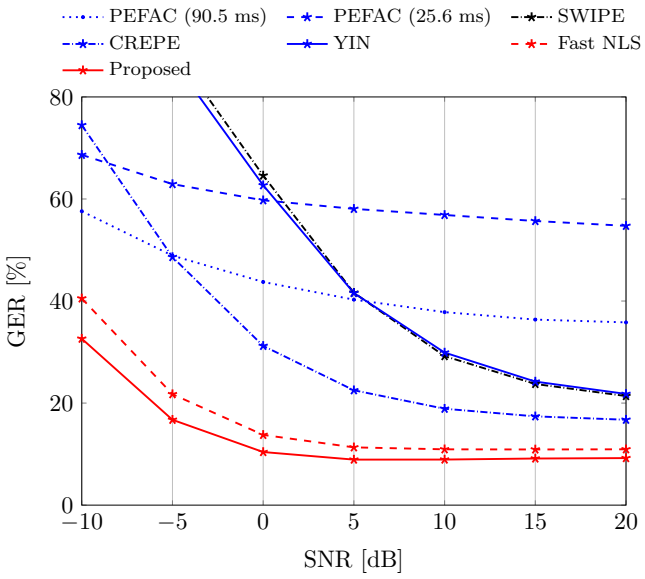

Fig. 7. Gross error ratio in different SNRs for the Keele pitch database in white Gaussian noise

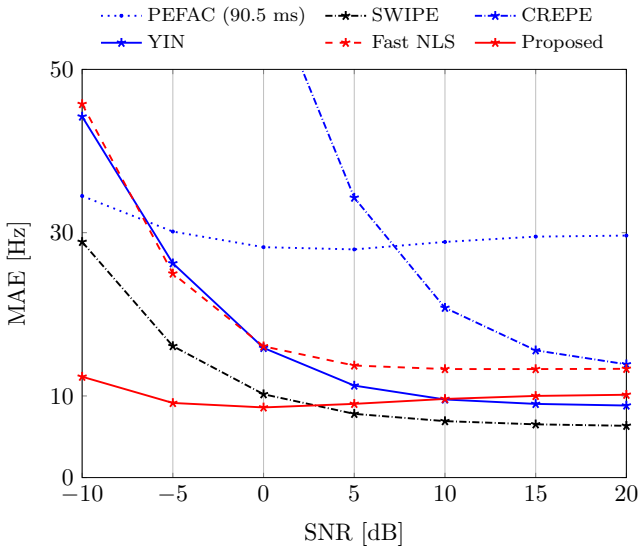

Fig. 8. Mean absolute error in different SNRs for the Keele pitch database with oracle voicing detector in white Gaussian noise

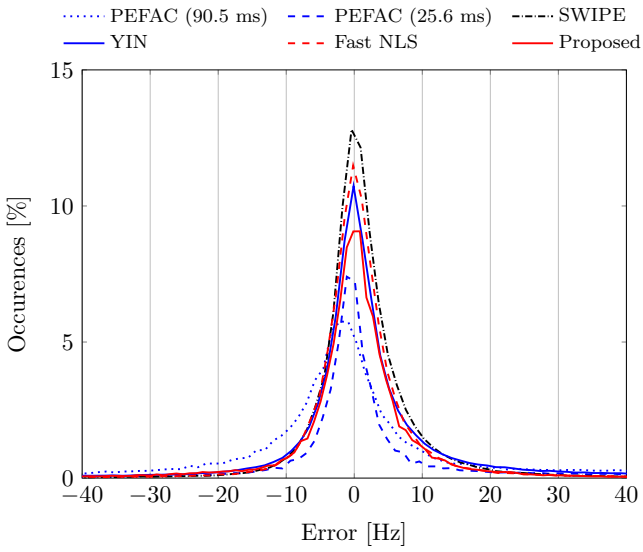

Fig. 9. Pitch estimation error distributions of different algorithms for the Keele pitch database with oracle voicing detector in -5 dB white Gaussian noise

GER and MAE in different SNRs for PEFAC, SWIPE, YIN, CREPE, fast NLS and the proposed algorithm are shown in Fig. 6, Fig. 7 and Fig. 8, respectively. The error distributions of PEFAC, SWIPE, YIN, Fast NLS and the proposed algorithm with oracle voicing detector in -5 dB white Gaussian noise are shown in Fig. 9. For YIN, fast NLS and the proposed algorithm, the frame length is set to the same as the reference, i.e., 25.6 ms. Frame lengths 25.6 ms and 90.5 ms (default value) are used for PEFAC. The other parameters are set to the same as for Fig. 4. Averages over 20 independent Monte Carlo experiments are used to compute TER, GER and MAE. The confidence intervals for them are not shown because they are not on the same scale as the mean values. For example, the 95% confidence intervals for GER and MAE estimates are on a scale of 0.1% and 0.1 Hz, respectively. As can be seen from Fig. 6 and Fig. 7, PEFAC has better performance in terms of both GER and TER than CREPE at -10 dB SNR. Moreover, using a longer frame length (90.5 ms) for PEFAC leads to a lower GER but a higher TER compared with a shorter frame length (25.6 ms). SWIPE and YIN have similar performance in terms of TER and GER. The fast NLS method outperforms the PEFAC, SWIPE, YIN and CREPE. By imposing a smoothing prior on the pitches, harmonic orders and the voicing and using the harmonic model combined, the proposed algorithm achieves lower GER and TER than the fast NLS. As can be seen from Fig. 8, when the oracle voicing detector is used, the SWIPE has the lowest MAE from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

TABLE I
TOTAL ERROR RATIO IN COLORED NOISE

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	0.42	0.38	0.34	0.29
	Factory	0.34	0.30	0.27	0.25
PEFAC (25.6 ms)	Babble	0.41	0.35	0.29	0.24
	Factory	0.30	0.25	0.22	0.21
SWIPE	Babble	0.50	0.42	0.29	0.19
	Factory	0.52	0.49	0.40	0.28
CREPE	Babble	0.40	0.29	0.21	0.16
	Factory	0.39	0.28	0.20	0.15
YIN	Babble	0.50	0.43	0.32	0.22
	Factory	0.50	0.43	0.32	0.22
Prew-Fast NLS	Babble	0.35	0.27	0.18	0.12
	Factory	0.28	0.20	**0.14**	**0.11**
Prew-Proposed	Babble	**0.34**	**0.25**	**0.17**	**0.12**
	Factory	**0.28**	**0.20**	0.15	0.12

TABLE II
GROSS ERROR RATIO IN COLORED NOISE

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	0.62	0.51	0.44	0.39
	Factory	0.56	0.47	0.41	0.38
PEFAC (25.6 ms)	Babble	0.72	0.65	0.60	0.57
	Factory	0.68	0.61	0.57	0.54
SWIPE	Babble	0.96	0.81	0.55	0.36
	Factory	1.00	0.94	0.76	0.54
CREPE	Babble	0.73	0.50	0.34	0.24
	Factory	0.75	0.53	0.36	0.26
YIN	Babble	0.95	0.83	0.61	0.42
	Factory	0.96	0.83	0.61	0.42
Prew-Fast NLS	Babble	0.57	0.41	0.30	0.24
	Factory	0.55	0.42	0.33	0.28
Prew-Proposed	Babble	**0.53**	**0.36**	**0.27**	**0.24**
	Factory	**0.51**	**0.37**	**0.29**	**0.25**

5 to 20 dB while the proposed algorithm achieves the best performance from -10 to 0 dB. From Fig. 9, we can conclude that, for pitch estimation errors in the range $[-40, 40]$ Hz, the error distributions of SWIPE, PEFAC (25.6 ms), Fast NLS and the proposed algorithm in -5 dB white Gaussian noise are approximately symmetric around zero, while PEFAC (90.5 ms) tends to underestimate the pitch.

Second, the performance of the proposed algorithm with prewhitening is tested on the Keele pitch database in colored noise conditions, i.e., babble noise [4] and factory noise [5]. The time durations of these two files are both above 60 s. In each Monte Carlo trial, a randomly selected segment of the noise signals, according to the length of the speech signals, are scaled based on the desired SNR and added to the speech signals to simulate colored, noisy signals. The TER, GER and MAE results for Prew-proposed, Prew-fast NLS, PEFAC, Yin and SWIPE are shown in I, II and III, respectively. The linear prediction order for the prewhitening is set to 30. The maximum allowed harmonic order for the proposed and fast NLS is set to 30. The other parameters are set to the same as for Fig. 6. As can be seen from TABLE I and II, PEFAC with 90.5 ms and 25.6 ms have a lower TER and GER than YIN and SWIPE in -5 and 0 SNR conditions. The Prew-Proposed and Prew-Fast NLS have lower voicing detection errors and Gross errors than YIN, PEFAC and SWIPE in both babble and factory noise conditions. Although similar performance in term of TER can be seen for Prew-Proposed and Prew-Fast NLS, the Prew-Proposed has a lower GER than Prew-Fast NLS. As can be seen from TABLE III, when the oracle voicing detector is used, the SWIPE achieves the lowest MAE in babble noise. The Prew-proposed has a comparable performance with the SWIPE in babble noise and has the best performance in factory noise.

Third, we investigate the effect of reverberation on the performance of different pitch estimation algorithms. Reverberation is the process of multi-path propagation and occurs when the speech or audio signals are recorded in an acoustically enclosed space. A commonly used metric to measure the reverberation is the reverberation time (RT60) [47]. The reverberated signals used for testing are generated by filtering the signal by synthetic room impulse responses (RIRs) with RT60 varying from 0.2 to 1 s in 0.1 s step. The dimension of the room is set to $10 \times 6 \times 4$ m. The distance between the source and microphone is set to 1 m. The RIRs are generated using the image method [48] and implemented using the RIR Generator toolbox [49]. The position of the receiver is fixed while the position of the source is varied randomly from 60 degrees left of the receiver to 60 degrees right of the receiver for each Monte Carlo experiment. The TER, GER and MAE results on the Keele pitch database for the proposed, fast NLS, PEFAC, Yin and SWIPE are shown in Fig. 10, Fig. 11 and Fig. 12, respectively, where the parameters are set to the same as for Fig. 6. As can be seen from Fig. 10, the PEFAC (90.5 ms) has the lowest voicing detection errors in more reverberated conditions (RT60 from 0.5 to 1 s) while the

TABLE III
MEAN ABSOLUTE VALUE [HZ] IN COLORED NOISE WITH ORACLE VOICING DETECTOR

SNR		-5.00	0.00	5.00	10.00
PEFAC (90.5 ms)	Babble	49.81	39.15	31.73	27.96
	Factory	36.20	31.24	27.97	26.69
PEFAC (25.6 ms)	Babble	81.49	72.65	65.71	60.54
	Factory	72.61	64.93	57.93	54.20
SWIPE	Babble	**31.73**	**17.94**	**10.95**	**8.04**
	Factory	43.91	27.02	16.02	10.51
CREPE	Babble	68.95	44.93	30.57	21.89
	Factory	79.00	52.41	34.51	24.70
YIN	Babble	56.25	39.05	23.86	14.96
	Factory	57.37	38.53	23.41	14.97
Prew-Fast NLS	Babble	64.81	45.79	31.45	23.79
	Factory	74.58	57.88	44.93	36.50
Prew-Proposed	Babble	33.33	17.91	12.22	10.81
	Factory	**19.32**	**13.20**	**11.23**	**10.48**

[4]Crowd Talking 1 file in https://www.soundjay.com/crowd-talking-1.html
[5]Factory Floor Noise 2 file in http://spib.linse.ufsc.br/noise.html

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

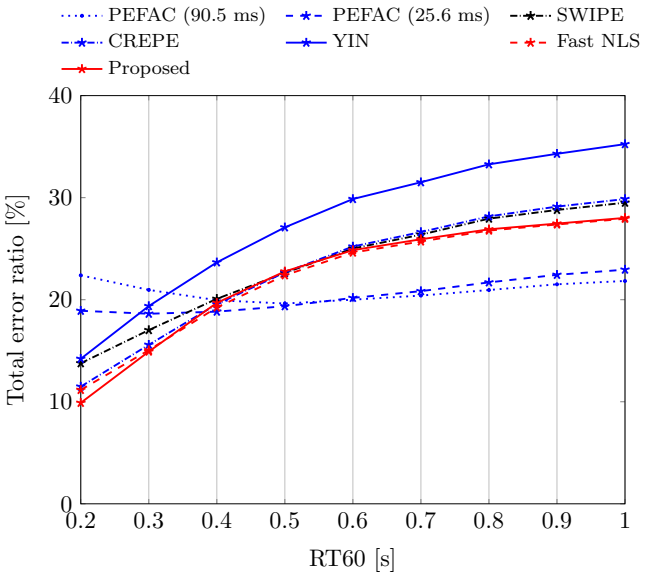

Fig. 10. Total error ratio in different reverberation time for the Keele pitch database

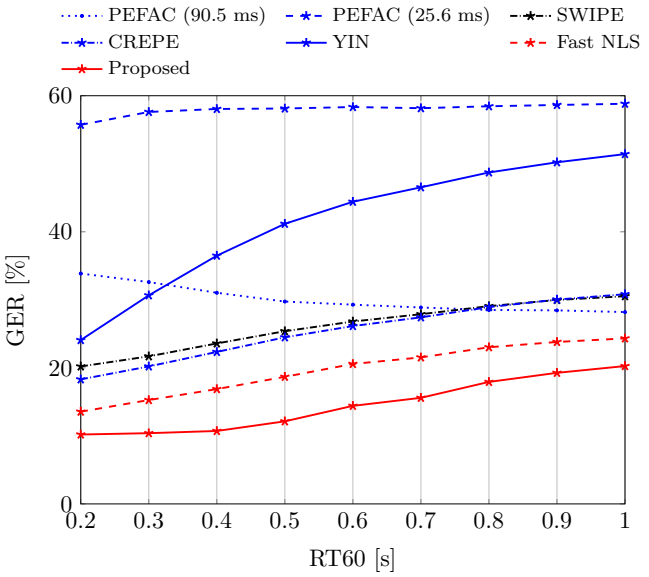

Fig. 11. Gross error ratio in different reverberation time for the Keele pitch database

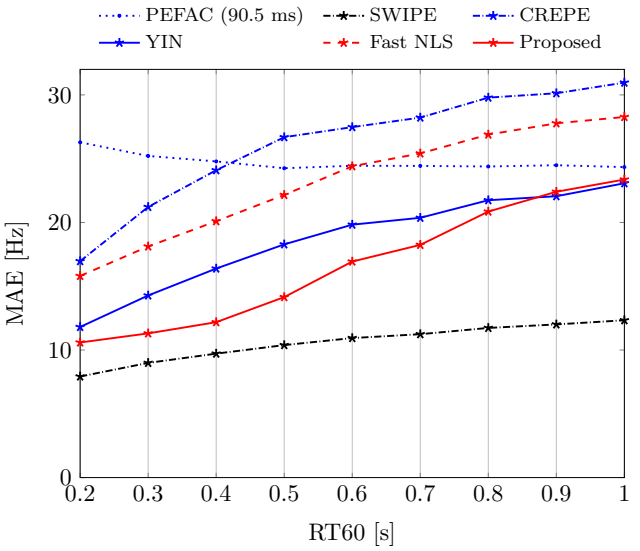

Fig. 12. Mean absolute error in different reverberation time for the Keele pitch database with oracle voicing detector

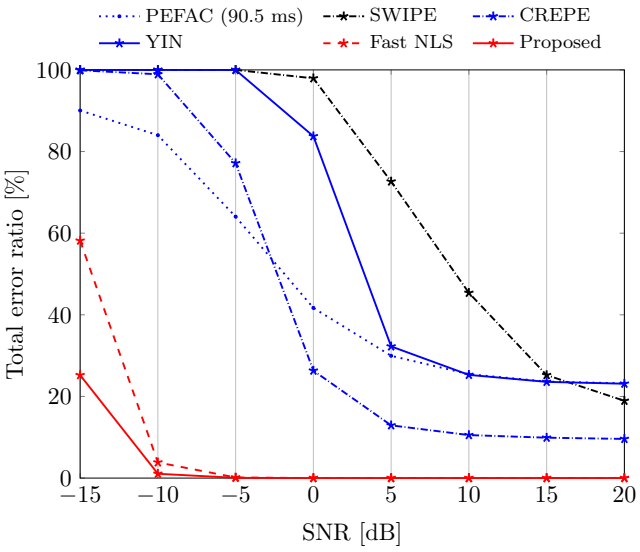

Fig. 13. Total error ratio in different SNRs for the Parkinson's disease database in white Gaussian noise

proposed algorithm has a better voicing detection performance in less reverberated conditions. The proposed and the fast NLS has similar performance in terms of TER. However, as can be seen from Fig. 11, the proposed outperforms the PEFAC, SWIPE, CREPE, YIN and fast NLS in terms of GER. From Fig. 12, we can conclude that SWIPE has the best performance while the proposed is the second best one in terms of MAE.

E. Experimental results on the Parkinson's disease database

In this subsection, the experimental results of different pitch estimation algorithms, using the Parkinson's disease database, in white Gaussian noise, is presented.

In the final experiment, the performance of the proposed algorithm is tested on sustained /a/ signals (voiced) from the Parkinson's disease database. The signals are downsampled to 16 kHz. TER, GER and MAE for different SNRs are shown in Fig. 13, Fig. 14 and Fig. 15, respectively. The error distributions of PEFAC, SWIPE, YIN, FAST NLS and the proposed algorithm with oracle voicing detector in -5 dB white Gaussian noise are shown in Fig. 16. The frame length is set to 80 ms for the fast NLS and proposed algorithms. The other parameters are set to the same as for Fig. 6. Similar conclusions to Fig. 6 and Fig. 7 can be drawn from Fig. 13 and Fig. 14. The proposed algorithm has the best performance in terms of the TER and GER. Moreover, the proposed algorithm has similar performance as SWIPE in terms of MAE measure from 5 to 20 dB and presents the lowest MAE from -15 to 0 dB. As can be seen from Fig. 16, for the Parkinson's disease database, the error distributions of PEFAC, SWIPE, Fast NLS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TASLP.2019.2930917, IEEE/ACM Transactions on Audio, Speech, and Language Processing

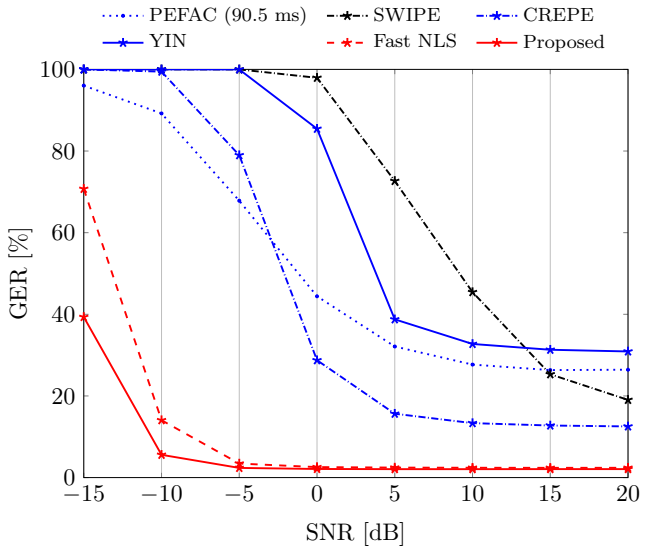

Fig. 14. Gross error ratio in different SNRs for the Parkinson's disease database in white Gaussian noise

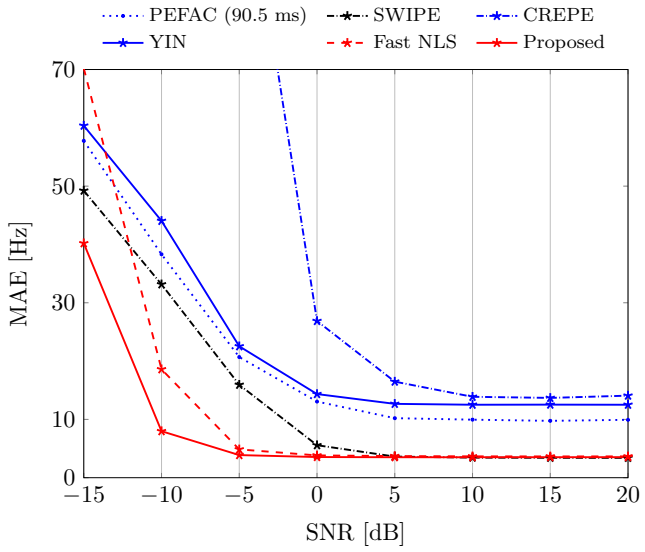

Fig. 15. Mean absolute error in different SNRs for the Parkinson's disease database with oracle voicing detector in white Gaussian noise

and the proposed algorithm in -5 dB white Gaussian noise are all approximately symmetric around zero. The spectrogram of one of the sustained /a/ sounds from the Parkinson's disease database, pitch estimates of the PEFAC (oracle), YIN (oracle) and the proposed algorithm in 0 dB white Gaussian noise are shown in Fig. 17. The oracle voicing detector from the ground truth (all voiced) is used for both PEFAC and YIN. As can be seen from Fig. 17, the proposed algorithm outperforms the PEFAC (oracle) and YIN (oracle).

Based on the above experiments, PEFAC obtains a better pitch estimation and voicing detection performance than the neural network-based CREPE in low SNR scenarios. SWIPE offers good performance in terms of MAE in high SNRs. The proposed algorithm obtains superior performance in terms of GER, TER and MAE compared to PEFAC, SWIPE, YIN,

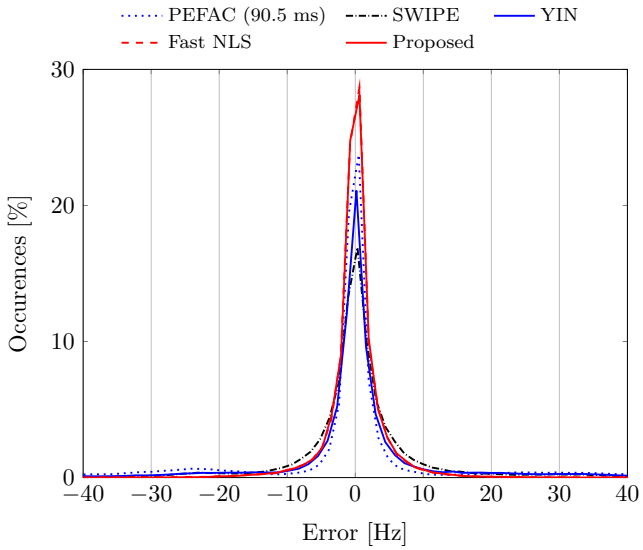

Fig. 16. Pitch estimation error distributions of different algorithms for the Parkinson's disease database with oracle voicing detector in -5 dB white Gaussian noise

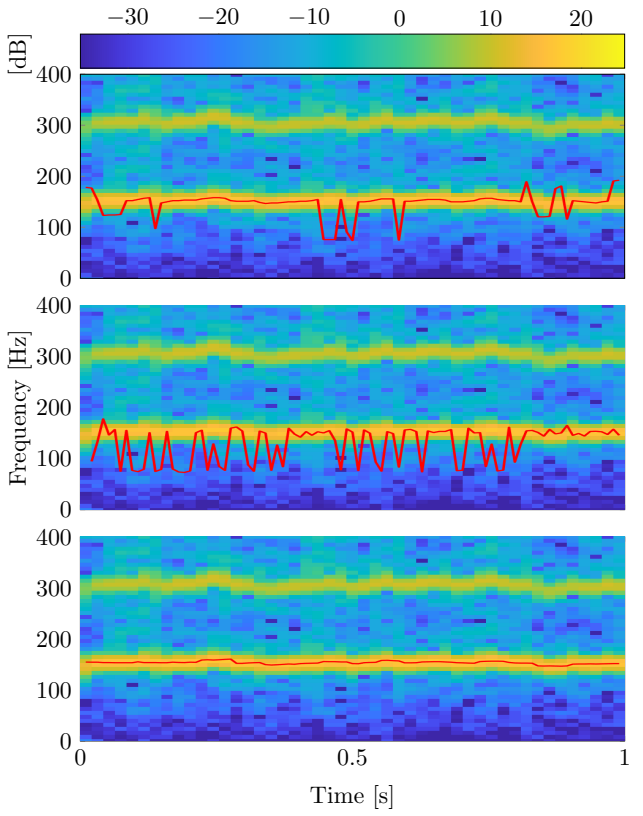

Fig. 17. Pitch estimates from PEFAC (oracle), YIN (oracle), and the proposed for sustained /a/ sounds from a database of Parkinson's disease voices in 0 dB white Gaussian noise.

CREPE, and the fast NLS in low SNR scenarios (under 5 dB) for the Keele pitch database and Parkinson's disease database. In high SNR scenarios (above 5 dB), the proposed algorithm has superior performance in terms of TER and GER, but not always the best performance in terms of MAE. In

practice, choosing pitch estimation algorithm depends on the applications and needs.

VIII. Conclusions

In this paper, a fully Bayesian harmonic model-based pitch tracking algorithm is proposed. Using a parametric harmonic model, the proposed algorithm shows good robustness against noise. The non-stationary evolution of the pitch, harmonic order and voicing state are modelled using first-order Markov chains. A fully Bayesian approach is applied for the noise variance and weights to avoid over-fitting. Using the hierarchical g-prior for the weights, the likelihood function can be easily evaluated using the fast NLS. The computational complexity of the recursive calculation of the predicted and posterior distributions is reduced by exploiting conditional independence between the pitch and harmonic order given the voicing indicators. Simulation results show that the proposed algorithm has good robustness against voicing state changes by carrying past information on pitch over the unvoiced/silent segments. The results of the pitch estimates and voicing detection for spoken sentences and sustained vowels are compared against ground truth estimates in the Keele and Parkinson's disease databases, showing that the proposed algorithm presents good pitch estimation and voicing detection accuracy even in very noisy conditions (e.g., -15 dB).

References

[1] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 184–194, April 2014.

[2] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinson's disease progression by noninvasive speech tests," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 884–893, 2010.

[3] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A pitch extraction algorithm tuned for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 2494–2498.

[4] D. Gerhard, *Pitch extraction and fundamental frequency: History and current techniques*. Department of Computer Science, University of Regina Regina, Canada, 2003.

[5] S. Gonzalez and M. Brookes, "PEFAC-a pitch estimation algorithm robust to high levels of noise," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 2, pp. 518–530, 2014.

[6] M. G. Christensen and A. Jakobsson, "Multi-pitch estimation," *Synthesis Lectures on Speech & Audio Processing*, vol. 5, no. 1, pp. 1–160, 2009.

[7] M. G. Christensen, "Accurate estimation of low fundamental frequencies from real-valued measurements," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 21, no. 10, pp. 2042–2056, Oct 2013.

[8] K. Paliwal and K. Wójcicki, "Effect of analysis window duration on speech intelligibility," *IEEE Signal Process. Lett.*, vol. 15, pp. 785–788, 2008.

[9] S. M. Nørholm, J. R. Jensen, and M. G. Christensen, "Instantaneous fundamental frequency estimation with optimal segmentation for non-stationary voiced speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 12, pp. 2354–2367, 2016.

[10] A. De Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[12] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, no. 3, pp. 1638–1652, 2008.

[13] J. K. Nielsen, T. L. Jensen, J. R. Jensen, M. G. Christensen, and S. H. Jensen, "Fast fundamental frequency estimation: Making a statistically efficient estimator computationally efficient," *Signal Process.*, vol. 135, pp. 188–197, 2017.

[14] G. Aneeja and B. Yegnanarayana, "Extraction of fundamental frequency from degraded speech using temporal envelopes at high SNR frequencies," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 25, no. 4, pp. 829–838, 2017.

[15] B. Quinn and P. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[16] J. Sward, H. Li, and A. Jakobsson, "Off-grid fundamental frequency estimation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 26, no. 2, pp. 296–303, Feb. 2018.

[17] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[18] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "A kalman-based fundamental frequency estimation algorithm," in *Proc. IEEE Workshop Appl. of Signal Process. to Aud. and Acoust.* IEEE Press, 2017, pp. 314–318.

[19] G. Zhang and S. Godsill, "Fundamental frequency estimation in speech signals with variable rate particle filters," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 5, pp. 890–900, 2016.

[20] C. Andrieu, M. Davy, and A. Doucet, "Efficient particle filtering for jump markov systems. application to time-varying autoregressions," *IEEE Trans. Signal Process.*, vol. 51, no. 7, pp. 1762–1770, 2003.

[21] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 1, pp. 76–87, 2004.

[22] E. Fisher, J. Tabrikian, and S. Dubnov, "Generalized likelihood ratio test for voiced-unvoiced decision in noisy speech using the harmonic model," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 14, no. 2, pp. 502–510, 2006.

[23] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[24] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 22, no. 12, pp. 2158–2168, 2014.

[25] X. Zhang, H. Zhang, S. Nie, G. Gao, and W. Liu, "A pairwise algorithm using the deep stacking network for speech separation and pitch estimation," *IEEE/ACM Trans. Audio, Speech, and Lang. Process.*, vol. 24, no. 6, pp. 1066–1078, 2016.

[26] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, April 2018, pp. 161–165.

[27] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 174–188, 2002.

[28] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[29] M. G. Christensen, P. Stoica, A. Jakobsson, and S. H. Jensen, "Multi-pitch estimation," *Signal Process.*, vol. 88, no. 4, pp. 972–983, 2008.

[30] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the itakura-saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793–830, 2009.

[31] O. Cappé, S. J. Godsill, and E. Moulines, "An overview of existing methods and recent advances in sequential monte carlo," *Proc. IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[32] W. Fong, S. J. Godsill, A. Doucet, and M. West, "Monte carlo smoothing with application to audio signal enhancement," *IEEE Trans. Signal Process.*, vol. 50, no. 2, pp. 438–449, 2002.

[33] A. Doucet, N. De Freitas, K. Murphy, and S. Russell, "Rao-blackwellised particle filtering for dynamic bayesian networks," in *Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2000, pp. 176–183.

[34] S. I. Adalbjörnsson, A. Jakobsson, and M. G. Christensen, "Multi-pitch estimation exploiting block sparsity," *Signal Process.*, vol. 109, pp. 236–247, 2015.

[35] J. K. Nielsen, M. G. Christensen, A. T. Cemgil, and S. H. Jensen, "Bayesian model comparison with the g-prior," *IEEE Trans. Signal Process.*, vol. 62, no. 1, pp. 225–238, 2014.

[36] A. Zellner, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian inference and decision techniques*, 1986.

[37] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for bayesian variable selection," *J. Amer. Stat. Assoc.*, vol. 103, no. 481, pp. 410–423, 2008.

[38] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, 1999.

[39] M. G. Christensen, A. Jakobsson, and S. H. Jensen, "Joint high-resolution fundamental frequency and order estimation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, no. 5, pp. 1635–1644, 2007.

[40] P. Stoica and Y. Selen, "Model-order selection: a review of information criterion rules," *IEEE Signal Process. Mag.*, vol. 21, no. 4, pp. 36–47, 2004.

[41] K. Yoshii and M. Goto, "A nonparametric bayesian multipitch analyzer based on infinite latent harmonic allocation," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 3, pp. 717–730, 2012.

[42] I. S. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*. Academic press, 2014.

[43] A. E. Jaramillo, J. K. Nielsen, and M. G. Christensen, "A study on how pre-whitening influences fundamental frequency estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 2019, pp. 6495–6499.

[44] T. Gerkmann and R. C. Hendriks, "Unbiased mmse-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[45] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in *Fourth European Conference on Speech Communication and Technology*, 1995.

[46] A. Tsanas, M. Zañartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford, "Robust fundamental frequency estimation in sustained vowels: detailed algorithmic comparisons and information fusion with adaptive kalman filtering," *J. Acoust. Soc. Am.*, vol. 135, no. 5, pp. 2885–2901, 2014.

[47] P. A. Naylor and N. D. Gaubitch, *Speech dereverberation*. Springer Science & Business Media, 2010.

[48] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[49] E. A. Habets, "Room impulse response generator," *Technische Universiteit Eindhoven, Tech. Rep*, vol. 2, no. 2.4, p. 1, 2006.